Plan

Today

- Wrap-up regularization (for now)
- MLE
  - Probability / likelihood
  - Maximum likelihood estimation
  - Probabilistic formulation of linear and logistic regression

Wrap up Neural Nets

Switch to regression slides

10-315
Introduction to ML

MLE and Probabilistic Formulation of Machine Learning

Instructor: Pat Virtue

Poll 1: Exercise

Implement a function in Python for the pdf of a Gaussian distribution.

Python numpy or math packages are fine, no scipy, etc.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Exercises

Calculate the probability of these event sequences happening

Coin

Fair: H, H, T, H

Biased, $\phi = 3/4$ heads: H, H, T, H

4-sided die with sides: A, B, C, D

Fair: A, B, D, D, A

Weighted, $[\phi_A, \phi_B, \phi_C, \phi_D] = [1/10,\ \ 2/10,\ \ 3/10,\ \ 4/10]$

A, B, D, D, A

Probability

Probability Vocab

Outcomes

Sample space

Events

Probability

Random variable

Discrete random variable

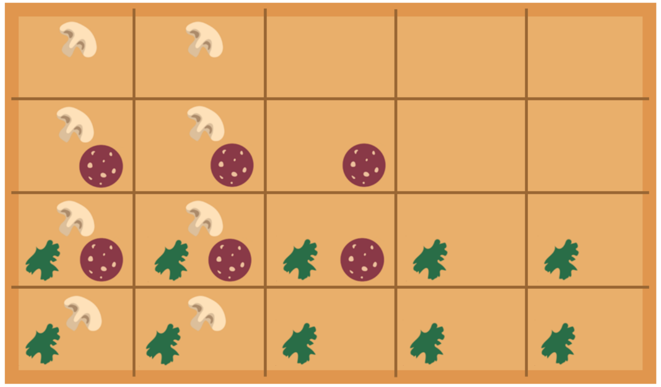
Continuous random variable

Probability mass function

Probability density function

Parameters

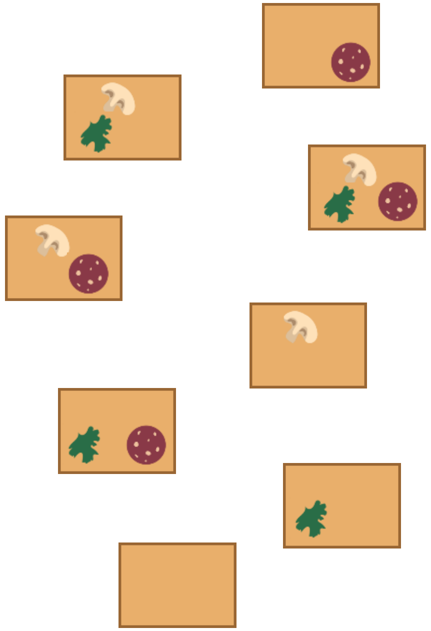
Example



Random variable for spinach or no spinach

Random variable

Distribution

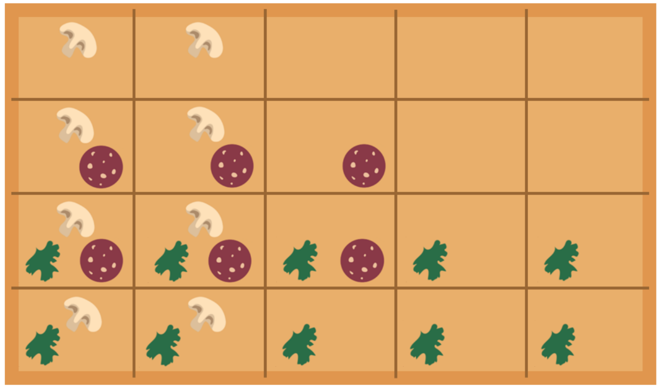


Sample space → Values → Probabilities or Densities

Icons: CC, https://openclipart.org/detail/296791/pizza-slice

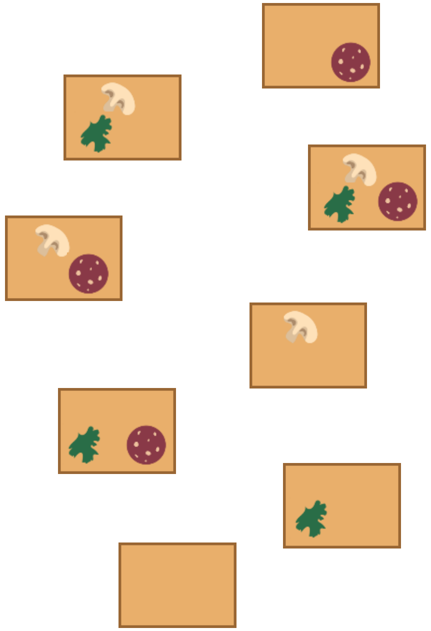
Example

Random variable for spinach or no spinach



Random variable

Distribution

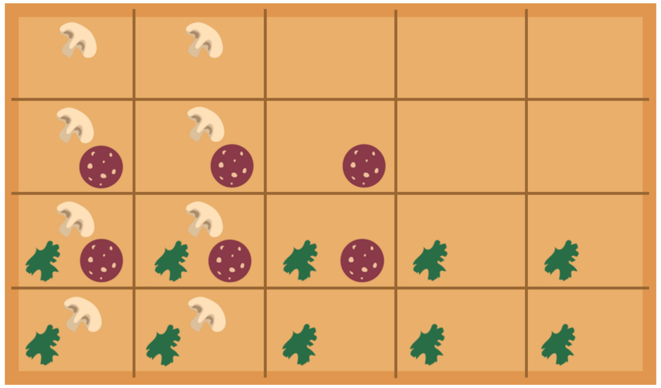


Sample space → Values → Probabilities or Densities

Icons: CC, https://openclipart.org/detail/296791/pizza-slice

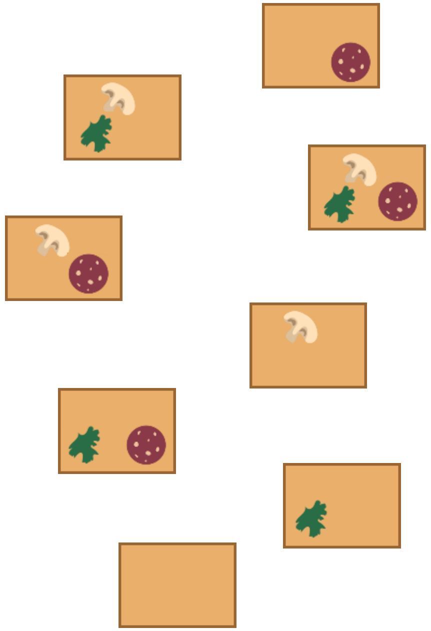
Example

Random variable for topping type with three categories: none, non-meat, meat



Random variable

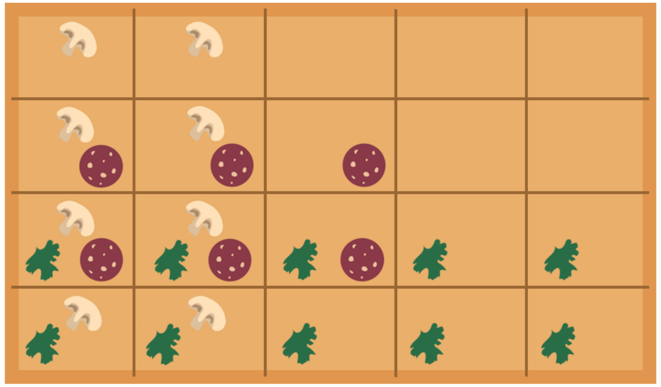
Distribution



Sample space → Values → Probabilities or Densities

Icons: CC, https://openclipart.org/detail/296791/pizza-slice

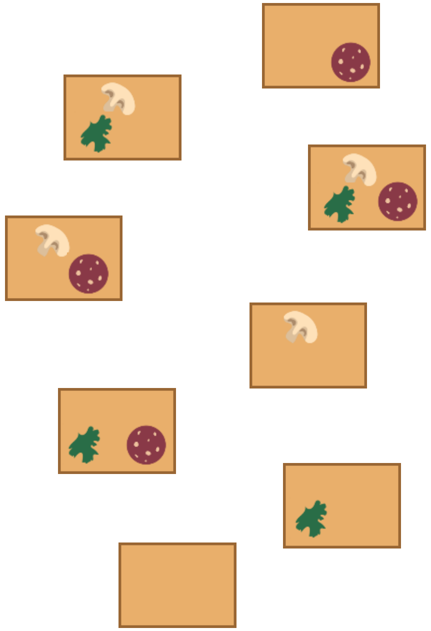
Example



Random variable for topping type with three categories: none, non-meat, meat

Random variable

Distribution



Sample space → Values → Probabilities or Densities

Icons: CC, https://openclipart.org/detail/296791/pizza-slice

Example

Random variable for number of heads after two flips of a fair coin

Random variable

Distribution

Sample space

Values

Probabilities or Densities

Example

Random variable for number of heads after two flips of a *biased* coin that lands heads 75%

Random variable | Distribution

Sample space → Values → Probabilities or Densities

Icons: CC, https://openclipart.org/detail/296791/pizza-slice

Example

Random variable for cat in picture or not

Random variable

Distribution

Sample space → Values → Probabilities or Densities

Example

Random variable for animal species in picture assuming one animal picture and available species: dog, cat, pig

Random variable — Distribution

Sample space → Values → Probabilities or Densities

Example

Random variable for height of student

Random variable

Distribution

Sample space → Values → Probabilities or Densities

Icons: CC, https://openclipart.org/detail/296791/pizza-slice

Probability Vocab

Outcomes

Sample space

Events

Probability

Random variable

Discrete random variable

Continuous random variable

Probability mass function

Probability density function

Parameters

Example Discrete Distributions

Bernoulli

Categorical

Binomial

Multinomial

Uniform

Example Continuous Distributions

Gaussian

Beta

Laplace

Probability Vocab

Marginal

Joint

Conditional

Notation

Dataset

Parameters, generically θ

$p(\mathcal{D} \mid \theta), p(\mathcal{D} \,;\, \theta)$

Random variables

Capital

Values

lower case

Random variable: function that maps events to values

Y is rand variable that maps the event of a coin toss being heads to value one and the event of a coin toss being tails to zero

$P(Y = 1 \mid \phi) = 3/4$, where $\phi = 3/4$

$P(Y = 1) = 3/4$

Sometimes even

$P(Y = heads) = 3/4$

Probability Toolbox

- Algebra
- Three axioms of probability
- Theorem of total probability
- Definition of conditional probability
- Product rule
- Bayes' theorem
- Chain rule
- Independence
- Conditional independence

Probability Tools Summary

Adding to the toolbox

1. Definition of conditional probability

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

2. Chain Rule

$$P(A,B) = P(A \mid B)P(B)$$

3. Bayes' theorem

$$P(B|A) = \frac{P(A \mid B)P(B)}{P(A)}$$

4. Chain Rule...

$$P(A_1, \dots A_N) = P(A_1) \sum_{i=2}^{N} P(A_i \mid A_{i-1})$$

Likelihood

Likelihood

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

Likelihood

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

Grades

Gaussian PDF: $p\left(y \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$

Likelihood

Trick coin: comes up heads only 1/3 of the time

1 flip: H probability: $\frac{1}{3}$

2 flips: H,H probability: $\frac{1}{3} \cdot \frac{1}{3}$

3 flips: H,H,T probability: $\frac{1}{3} \cdot \frac{1}{3} \cdot \left(1 - \frac{1}{3}\right)$

But why can we just multiply these?

Likelihood and i.i.d

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

i.i.d.: Independent and identically distributed

Bernoulli Likelihood

Bernoulli distribution:

$$Y \sim Bern(\phi) \qquad p(y \mid \phi) = \begin{cases} \phi, & y = 1 \\ 1 - \phi, & y = 0 \end{cases}$$

What is the likelihood for three i.i.d. samples, given parameter ϕ:

$$\mathcal{D} = \{y^{(1)} = 1, y^{(2)} = 1, y^{(3)} = 0\}$$

$$\prod_{i=1}^{N} p(Y = y^{(i)} \mid \phi)$$

$$= \phi \cdot \phi \cdot (1 - \phi)$$

MLE

Maximum likelihood estimation

From Probability to Statistics

Poll 2

Assume that exam scores are drawn independently from the same Gaussian (Normal) distribution.

Given three exam scores 75, 80, 90, which pair of parameters is a better fit (a higher likelihood)?

A) Mean 80, standard deviation 3

B) Mean 85, standard deviation 7

Use a calculator/computer.

Gaussian PDF: $p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$

Likelihood

Trick coin

Estimating Parameters with Likelihood

We model the outcome of a single mysterious weighted-coin flip as a Bernoulli random variable:

$$Y \sim Bern(\phi)$$

$$p(y \mid \phi) = \begin{cases} \phi, & y = 1 \ \ (Heads) \\ 1 - \phi, & y = 0 \ \ (Tails) \end{cases}$$

Given the ordered sequence of coin flip outcomes:

$$[1, 0, 1, 1]$$

What is the estimate of parameter $\hat{\phi}$?

Estimating Parameters with Likelihood

We model the outcome of a single mysterious weighted-coin flip as a Bernoulli random variable:

$$Y \sim Bern(\phi)$$

$$p(y \mid \phi) = \begin{cases} \phi, & y = 1 \ \ (Heads) \\ 1 - \phi, & y = 0 \ \ (Tails) \end{cases}$$

Given the ordered sequence of coin flip outcomes:

$$[1, 0, 1, 1]$$

What is the estimate of parameter $\hat{\phi}$?

$$\begin{aligned} p(\, D \mid \phi \,) &= \phi \cdot \phi \cdot (1 - \phi) \cdot \phi \\ &= \phi^3 (1 - \phi)^1 \end{aligned}$$

Likelihood and Maximum Likelihood Estimation

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

Likelihood function: The value of likelihood as we change theta

(same as likelihood, but conceptually we are considering many different values of the parameters)

Maximum Likelihood Estimation (MLE): Find the parameter value that maximizes the likelihood.

MLE as Data Increases

Given the ordered sequence of coin flip outcomes:

$$[1, 0, 1, 1]$$

$$\mathrm{p}(\mathcal{D} \mid \phi) = \prod_{i}^{N} p\left(y^{(i)} \mid \phi \right) = \phi^{N_{y=1}} (1 - \phi)^{N_{y=0}}$$

What happens as we flip more coins?

MLE for Gaussian

Gaussian distribution:

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

What is the log likelihood for three i.i.d. samples, given parameters μ, σ^2?

$$\mathcal{D} = \{y^{(1)} = 65, y^{(2)} = 95, y^{(3)} = 85\}$$

$$L(\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y^{(i)}-\mu\right)^2}{2\sigma^2}}$$

$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} p\left(y^{(i)} \mid \boldsymbol{\theta}\right)$$

MLE for Gaussian

Assume that exam scores are drawn independently from the same Gaussian (Normal) distribution.

Given three exam scores 75, 80, 90, which pair of parameters is the best fit (the highest likelihood)?

$$p(\mathcal{D}|\mu,\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y^{(i)}-\mu\right)^2}{2\sigma^2}}$$

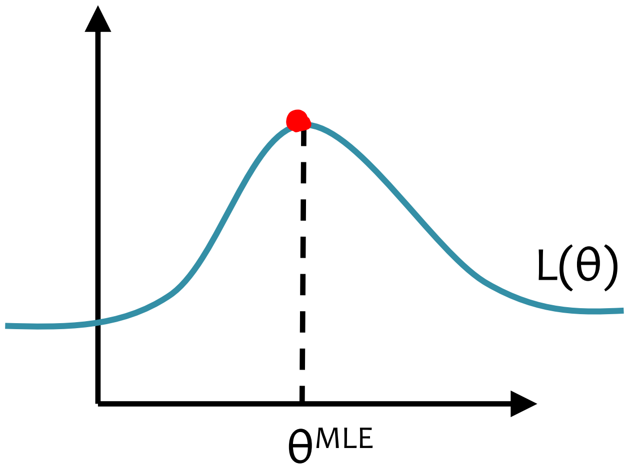
MLE

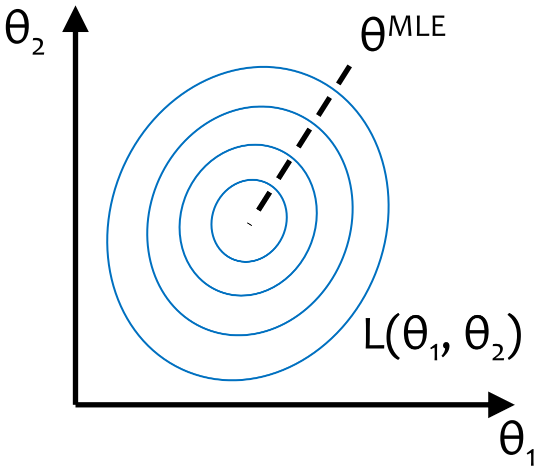
Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

Principle of Maximum Likelihood Estimation:
Choose the parameters that maximize the likelihood of the data.

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)} | \boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)





Slide credit: CMU MLD Matt Gormley

Likelihood and Log Likelihood

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

Likelihood function: The value of likelihood as we change theta

(same as likelihood, but conceptually we are considering many different values of the parameters)

Recipe for Estimation

MLE

1. Formulate the likelihood, $p(\mathcal{D} \mid \theta)$
2. Set objective $J(\theta)$ equal to negative log of likelihood

$$\mathrm{J}(\theta) = -\log p(\mathcal{D} \mid \theta)$$

3. Compute derivative of objective, $\partial J / \partial \theta$
4. Find $\hat{\theta}$, either
   a. Set derivate equal to zero and solve for θ
   b. Use (stochastic) gradient descent to step towards better θ

MLE for Gaussian

Gaussian distribution:

$$Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$p\left(y \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

What is the log likelihood for three i.i.d. samples, given parameters μ, σ^2?

$$\mathcal{D} = \{y^{(1)} = 65, y^{(2)} = 95, y^{(3)} = 85\}$$

$$L\left(\mu, \sigma^2\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y^{(i)}-\mu\right)^2}{2\sigma^2}}$$

$$\ell\left(\mu, \sigma^2\right) = \sum_{i=1}^{N} -\log\sqrt{2\pi\sigma^2} - \frac{\left(y^{(i)} - \mu\right)^2}{2\sigma^2}$$

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{i}^{N} p\left(y^{(i)} \mid \boldsymbol{\theta}\right)$$

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i}^{N} \log p\left(y^{(i)} \mid \boldsymbol{\theta}\right)$$

Probabilistic Formulation for ML

MLE for Linear and Logistic Regression

Using Statistics for Machine Learning

Recipe for Estimation

MLE

1. Formulate the likelihood, $p(\mathcal{D} \mid \theta)$
2. Set objective $J(\theta)$ equal to negative log of likelihood

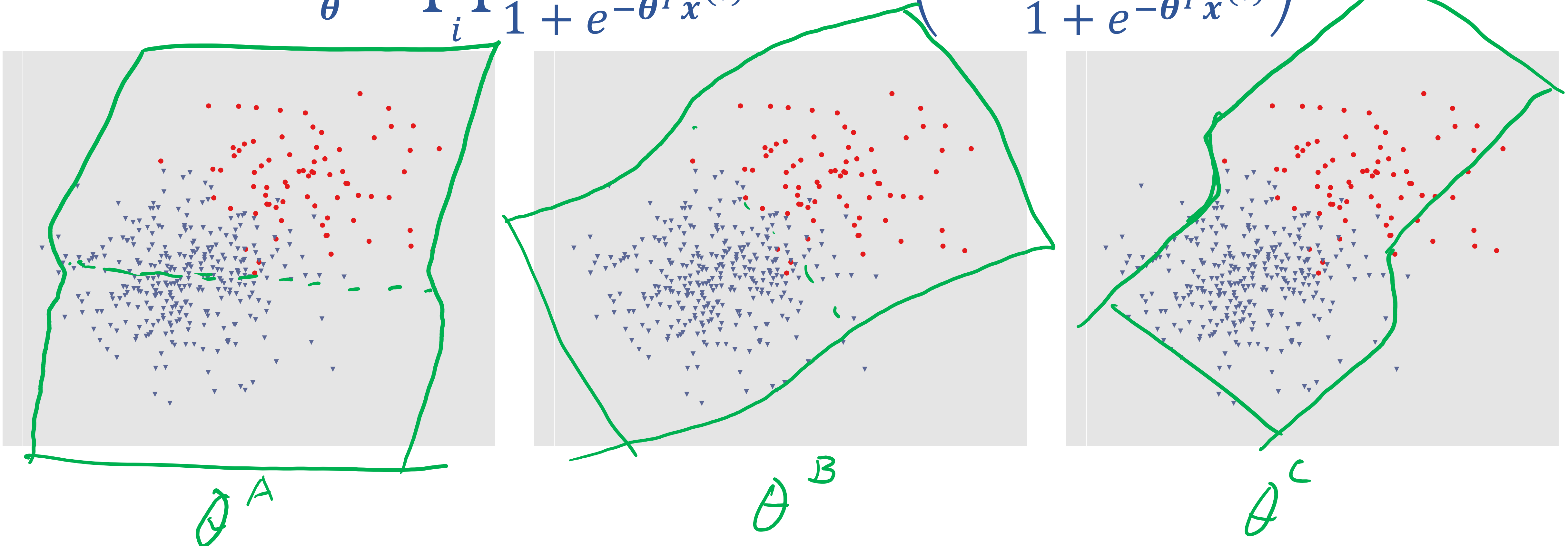
$$\mathrm{J}(\theta) = -\log p(\mathcal{D} \mid \theta)$$

3. Compute derivative of objective, $\partial J / \partial \theta$
4. Find $\hat{\theta}$, either
   a. Set derivate equal to zero and solve for θ
   b. Use (stochastic) gradient descent to step towards better θ

M(C)LE for Logistic Regression

Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of just one test results, X_A and X_B

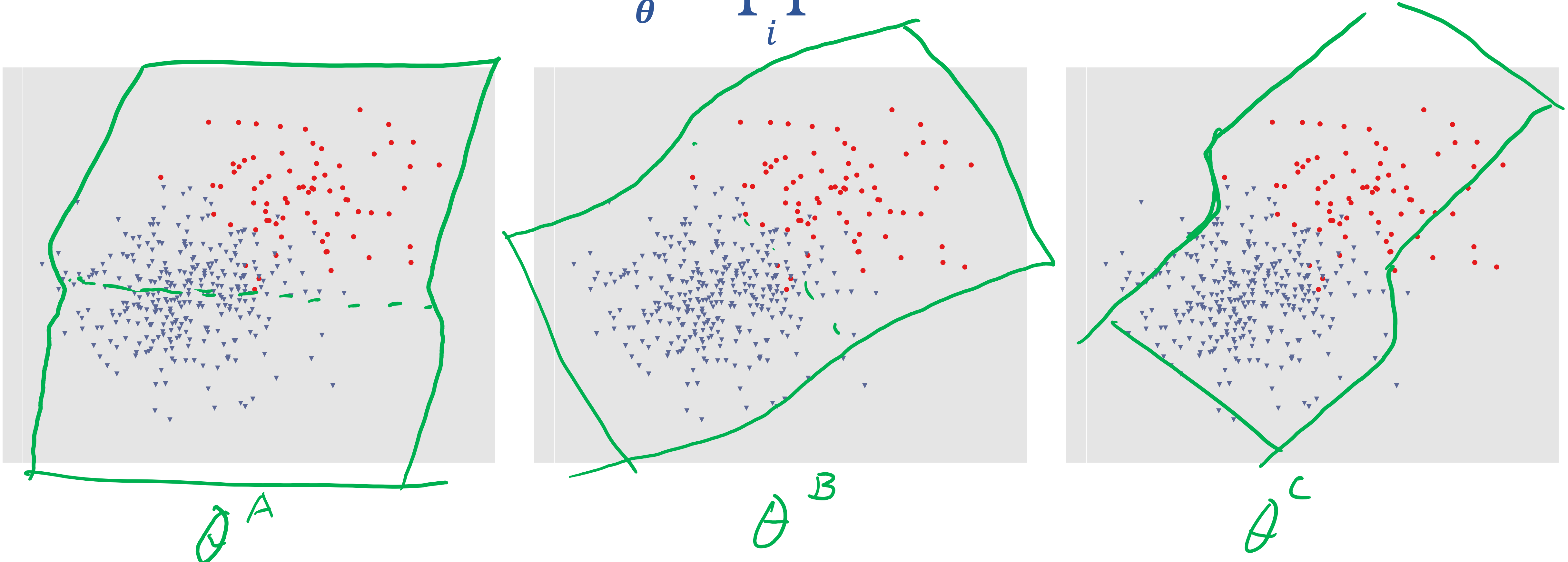
$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} \left(\frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}\right)^{\mathbb{I}(y^{(i)}=1)} \left(1 - \frac{1}{1 + e^{-\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}}\right)^{\mathbb{I}(y^{(i)}=0)}$$



M(C)LE for Logistic Regression

Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of just one test results, X_A and X_B

$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} p(\, y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\,)$$



M(C)LE for Multi-class Logistic Regression

Learn to predict if probability of output belonging to class k, Y_k , given input X, $P(Y_k = 1 \mid X, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$

$$\hat{\Theta}_{MLE} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \prod_{i}^{N} \prod_{k}^{K} \frac{e^{\boldsymbol{\theta}_k^T \boldsymbol{x}^{(i)}}}{\sum_{l=1}^{K} e^{\boldsymbol{\theta}_l^T \boldsymbol{x}^{(i)}}}^{\mathbb{I}\left(y_k^{(i)}=1\right)}$$

M(C)LE for Multi-class Logistic Regression

Learn to predict if probability of output belonging to class k, Y_k , given input X, $P(Y_k = 1 \mid X, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$

$$L(\Theta; \mathcal{D}) = \prod_{i}^{N} \prod_{k}^{K} \frac{e^{\boldsymbol{\theta}_k^T \boldsymbol{x}^{(i)}}}{\sum_{l=1}^{K} e^{\boldsymbol{\theta}_l^T \boldsymbol{x}^{(i)}}}^{\mathbb{I}\left(y_k^{(i)}=1\right)}$$

M(C)LE for Linear Regression

Probabilistic interpretation of linear regression

$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} p(\, y^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta} \,)$$

M(C)LE for Linear Regression

Probabilistic interpretation of linear regression

$$L(\theta; \mathcal{D}) = \prod_{i}^{N} p\left(y^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta} \right)$$